

## **NUCLEIC ACID MOLECULES AND OTHER MOLECULES ASSOCIATED WITH PLANTS AND USES THEREOF FOR PLANT IMPROVEMENT**

This application claims the benefit of and is a continuation in part of prior U.S.

5 application No. 09/684,016 filed October 10, 2000, and prior U.S. application No. 09/850,147  
filed May 7, 2001, both of which are hereby incorporated by reference in their entirety.

### **INCORPORATION OF SEQUENCE LISTING**

Two copies of the sequence listing (Seq. Listing Copy 1 and Seq. Listing Copy 2) and a  
computer-readable form of the sequence listing, all on CD-ROMs, each containing the file  
10 named pa\_00620.rpt, which is 74,252,288 bytes (measured in MS-DOS) and was created on  
January 20, 2004, are herein incorporated by reference.

### **INCORPORATION OF TABLE**

Two copies of Table 1 (Table 1 Copy 1 and Table 1 Copy 2) all on CD-ROMs, each  
containing the file named pa\_00620.txt, which is 8,415,232 bytes (measured in MS-DOS) and  
15 was created on January 20, 2004, are herein incorporated by reference.

### **FIELD OF THE INVENTION**

Disclosed herein are inventions in the field of plant biochemistry and genetics. More  
specifically recombinant polynucleotides and recombinant polypeptides from *Sorghum* for use in  
plant improvement are provided. Methods of using the recombinant polynucleotides and  
20 recombinant polypeptides for production of transgenic plants with improved biological  
characteristics are disclosed.

## BACKGROUND OF THE INVENTION

The ability to develop transgenic plants with improved traits depends in part on the identification of polynucleotides that are useful for the production of transformed plants having desirable qualities. In this regard, the discovery of polynucleotide sequences of genes, and the polypeptides encoded by such genes, is needed. Molecules comprising such polynucleotides may be used, for example, in recombinant DNA constructs useful for imparting unique genetic properties into transgenic plants.

## SUMMARY OF THE INVENTION

The present invention provides a recombinant polynucleotide selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 31,564. The present invention also provides a recombinant polypeptide selected from the group consisting of SEQ ID NO: 31,565 through SEQ ID NO: 63,128.

The present invention also provides a method of producing a plant having an improved property, wherein said method comprises transforming a plant with a recombinant construct comprising a promoter region functional in a plant cell operably joined to a polynucleotide comprising a coding sequence for a polypeptide associated with said property, and growing said transformed plant.

## DETAILED DESCRIPTION OF THE INVENTION

The present invention provides recombinant polynucleotides and recombinant polypeptides from *Sorghum*. The recombinant polynucleotides and recombinant polypeptides of the present invention find a number of uses, for example in recombinant DNA constructs, in

physical arrays of molecules, for use as plant breeding markers, and for use in computer based storage and analysis systems.

The recombinant polynucleotides of the present invention also find use in generation of transgenic plants to provide for increased or decreased expression of the polypeptides encoded by the recombinant polynucleotides provided herein. As used herein a “transgenic” organism is one whose genome has been altered by the incorporation of foreign genetic material or additional copies of native genetic material, e.g. by transformation or recombination. As a result of such biotechnological applications, plants, particularly crop plants, having improved properties are obtained. Crop plants of interest in the present invention include, but are not limited to soy, cotton, canola, maize, wheat, sunflower, sorghum, alfalfa, barley, millet, rice, tobacco, fruit and vegetable crops, and turf grass. In one embodiment the disclosed recombinant polynucleotides provide plants having improved yield resulting from improved utilization of key biochemical compounds, such as nitrogen, phosphorous and carbohydrate, or resulting from improved responses to environmental stresses, such as cold, heat, drought, salt, and attack by pests or pathogens. Recombinant polynucleotides of the present invention may be used to provide plants having improved growth and development, and ultimately increased yield, as the result of modified expression of plant growth regulators or modification of cell cycle or photosynthesis pathways. Other traits of interest that may be modified in plants using polynucleotides of the present invention include flavonoid content, seed oil and protein quantity and quality, herbicide tolerance, and rate of homologous recombination.

### **Polynucleotides**

Depending on the intended use, the recombinant polynucleotides of the present invention may be present in the form of DNA, such as cDNA or genomic DNA, or as RNA, for example

mRNA. The polynucleotides of the present invention may be single or double stranded and may represent the coding, or sense strand of a gene, or the non-coding, antisense, strand. In one embodiment, the recombinant polynucleotides of this invention represent cDNA sequences from *Sorghum*. DNA sequences representing the recombinant polynucleotides are provided herein as

5 SEQ ID NO: 1 through SEQ ID NO: 31,564.

The term “recombinant polynucleotide” as used herein refers to a polynucleotide produced by recombinant DNA technology. In one embodiment a recombinant polynucleotide may be produced by separation from substantially all other molecules normally associated with it in its native state. A recombinant polynucleotide may be greater than 60% free, greater than  
10 75% free, greater than 90% free, or greater than 95% free from the other molecules (exclusive of solvent) present in the natural mixture. In another embodiment, a recombinant polynucleotide may be separated from nucleic acids which normally flank the polynucleotide in nature. Thus, polynucleotides fused to regulatory or coding sequences with which they are not normally associated, for example as the result of recombinant techniques, are considered recombinant  
15 polynucleotides herein. Such molecules are considered recombinant polynucleotides even when present, for example in the chromosome of a host cell, or in a nucleic acid solution. The term recombinant polynucleotide as used herein is not intended to encompass molecules present in their native state.

It is understood that the molecules of the invention may be labeled with reagents that  
20 facilitate detection of the molecule. As used herein, a label can be any reagent that facilitates detection, including fluorescent labels, chemical labels, or modified bases, including nucleotides with radioactive elements, e.g.  $^{32}\text{P}$ ,  $^{33}\text{P}$ ,  $^{35}\text{S}$  or  $^{125}\text{I}$  such as  $^{32}\text{P}$  deoxycytidine-5'-triphosphate ( $^{32}\text{PdCTP}$ ).

Recombinant polynucleotides of the present invention are capable of specifically hybridizing to other polynucleotides under certain circumstances. As used herein, two polynucleotides are said to be capable of specifically hybridizing to one another if the two molecules are capable of forming an anti-parallel, double-stranded nucleic acid structure. A polynucleotide is said to be the “complement” of another polynucleotide if the molecules exhibit complete complementarity. As used herein, molecules are said to exhibit “complete complementarity” when every nucleotide in each of the polynucleotides is complementary to the corresponding nucleotide of the other. Two polynucleotides are said to be “minimally complementary” if they can hybridize to one another with sufficient stability to permit them to remain annealed to one another under at least conventional “low-stringency” conditions. Similarly, the polynucleotides are said to be “complementary” if they can hybridize to one another with sufficient stability to permit them to remain annealed to one another under conventional “high-stringency” conditions. Conventional stringency conditions are known to those skilled in the art and can be found, for example in *Molecular Cloning: A Laboratory Manual*, 3<sup>rd</sup> edition Volumes 1, 2, and 3. J.F. Sambrook, D.W. Russell, and N. Irwin, Cold Spring Harbor Laboratory Press, 2000.

Departures from complete complementarity are therefore permissible, as long as such departures do not completely preclude the capacity of the polynucleotides to form a double-stranded structure. Thus, in order for a polynucleotide to serve as a primer or probe it need only be sufficiently complementary in sequence to be able to form a stable double-stranded structure under the particular solvent and salt concentrations employed. Appropriate stringency conditions which promote DNA hybridization are, for example, 6.0 X sodium chloride/sodium citrate (SSC) at about 45°C, followed by a wash of 2.0 X SSC at 50°C. Such conditions are known to those

skilled in the art and can be found, for example in *Current Protocols in Molecular Biology*, John Wiley & Sons, N.Y. (1989). Salt concentration and temperature in the wash step can be adjusted to alter hybridization stringency. For example, conditions may vary from low stringency of about 2.0 x SSC at 40°C to moderately stringent conditions of about 2.0 x SSC at 50°C to  
5 high stringency conditions of about 0.2 x SSC at 50°C.

As used herein “sequence identity” refers to the extent to which two optimally aligned polynucleotides or polypeptide sequences are invariant throughout a window of alignment of components, e.g. nucleotides or amino acids. An “identity fraction” for aligned segments of a test sequence and a reference sequence is the number of identical components which are shared  
10 by the two aligned sequences divided by the total number of components in the reference sequence segment, i.e. the entire reference sequence or a smaller defined part of the reference sequence. “Percent identity” is the identity fraction times 100. Comparison of sequences to determine percent identity can be accomplished by a number of well-known methods, including for example by using mathematical algorithms, such as those in the BLAST suite of sequence  
15 analysis programs.

In one embodiment this invention provides recombinant polynucleotides comprising regions that encode polypeptides. The encoded polypeptides may be the complete protein encoded by the gene represented by the polynucleotide, or may be fragments of the encoded protein. In one embodiment, polynucleotides provided herein encode polypeptides constituting  
20 a substantial portion of the complete protein. In another embodiment polynucleotides provided herein encode polypeptides constituting a sufficient portion of the complete protein to provide the relevant biological activity.

In one embodiment recombinant polynucleotides of the present invention encode polypeptides involved in one or more important biological function in plants. Such recombinant polynucleotides may be expressed in transgenic plants to produce plants having improved phenotypic properties and/or improved response to stressful environmental conditions. See, for example, Table 1 for a list of SEQ ID numbers representing the recombinant polynucleotides that may be expressed in transgenic plants to impart an improved plant property where improved plant properties are provided for each sequence in the PRODUCT\_CAT\_DESC column.

Recombinant polynucleotides of the present invention are generally used to impart such improved properties by providing for enhanced protein activity in a transgenic organism, such as a transgenic plant, although in some cases, improved properties are obtained by providing for reduced protein activity in a transgenic plant. Reduced protein activity and enhanced protein activity are measured by reference to a wild type cell or organism and can be determined by direct or indirect measurement. Direct measurement of protein activity might include an analytical assay for the protein, per se, or enzymatic product of protein activity. Indirect assay might include measurement of a property affected by the protein. Enhanced protein activity can be achieved in a number of ways, for example by overproduction of mRNA encoding the protein or by gene shuffling. One skilled in the art will know methods to achieve overproduction of mRNA, for example by providing increased recombinant copies of a gene or by introducing a recombinant construct having a heterologous promoter operably linked to a recombinant polynucleotide encoding a polypeptide into a target cell or organism. Reduced protein activity can be achieved by a variety of mechanisms including antisense, mutation, or knockout. Antisense RNA will reduce the level of expressed protein resulting in reduced protein activity as compared to wild type activity levels. A mutation in the gene encoding a protein may reduce the

level of expressed protein and/or interfere with the function of expressed protein to cause reduced protein activity.

In one embodiment, the invention is a fragment of a disclosed recombinant polynucleotide consisting of oligonucleotides of at least 15, at least 16 or 17, at least 18 or 19, or  
5 at least 20 or more consecutive nucleotides. Such oligonucleotides are fragments of the larger recombinant polynucleotides having a sequence selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 31,564, and find use, for example as probes and primers for detection of the polynucleotides of the present invention.

In one embodiment the present invention is a functional variant of a recombinant  
10 polynucleotide provided herein. As used herein, a “functional variant” refers to any second polynucleotide varying from a first polynucleotide sequence in such a way so as not to significantly affect the function when compared to the function of the first polynucleotide. Such functional variants may be naturally occurring, including homologous polynucleotides from the same or a different species, or may be non-natural functional variants, for example  
15 polynucleotides synthesized using chemical synthesis methods, or generated using recombinant DNA techniques. With respect to nucleotide sequences, degeneracy of the genetic code provides the possibility to substitute at least one base of the protein encoding sequence of a gene with a different base without causing the amino acid sequence of the polypeptide produced from the gene to be changed. Hence, in one embodiment, a recombinant polynucleotide of the present  
20 invention may have any base sequence that has been changed from SEQ ID NO: 1 through SEQ ID NO: 31,564 by substitution in accordance with degeneracy of the genetic code. See for example, US Patent number 5,500,365, which is hereby incorporated by reference.



Polynucleotides of the present invention that are functional variants of the polynucleotides provided herein will generally demonstrate significant identity with the polynucleotides provided herein. Of particular interest are polynucleotide homologs having at least about 60% sequence identity, at least about 70% sequence identity, at least about 80% sequence identity, at least about 85% sequence identity, and at least about 90%, 95% or even greater, such as 98% or 99% sequence identity with polynucleotide sequences described herein.

### **Polypeptides**

In one embodiment this invention also provides recombinant polypeptides. Amino acid sequences of the recombinant polypeptides of the present invention are provided herein as SEQ ID NO: 31,565 through SEQ ID NO: 63,128.

As used herein, the term “polypeptide” refers to an unbranched chain of amino acid residues that are covalently linked by an amide linkage between the carboxyl group of one amino acid and the amino group of another. The term polypeptide can encompass whole proteins (i.e. a functional protein encoded by a particular gene), as well as fragments of proteins. In one embodiment the invention is a recombinant polypeptide which represents a whole protein. In another embodiment the invention is a recombinant polypeptide which represents a sufficient portion of an entire protein to impart the relevant biological activity of the protein. The term “protein” also includes molecules consisting of one or more polypeptide chains. Thus, a recombinant polypeptide of the present invention may also constitute an entire gene product, but only a portion of a functional oligomeric protein having multiple polypeptide chains.

As used herein, the term “recombinant polypeptide” refers to a polypeptide produced as a result of recombinant DNA technology. The term recombinant polypeptide as used herein is not intended to encompass molecules present in their native state.

In one embodiment the invention is a recombinant polypeptide involved in one or more important biological properties in a plant. Such recombinant polypeptide may be produced in transgenic plants to provide plants having improved phenotypic properties and/or improved response to stressful environmental conditions. In some cases, decreased expression of such polypeptide may be desired, such decreased expression being obtained by use of the polynucleotide sequences provided herein, for example in antisense or cosuppression methods. See, Table 1 for a list of improved plant properties and PROTEIN\_NUM for the recombinant polypeptide whose expression may be altered in transgenic plants to impart such improvements. A summary of such improved properties and polypeptides of interest for increased or decreased expression is provided below.

Yield/Nitrogen: Yield improvement by improved nitrogen flow, sensing, uptake, storage and/or transport. Polypeptides useful for imparting such properties include those involved in aspartate and glutamate biosynthesis, polypeptides involved in aspartate and glutamate transport, polypeptides associated with the TOR (Target of Rapamycin) pathway, nitrate transporters, ammonium transporters, chlorate transporters and polypeptides involved in tetrapyrrole biosynthesis.

Yield/Carbohydrate: Yield improvement by effects on carbohydrate metabolism, for example by increased sucrose production and/or transport. Polypeptides useful for improved yield by effects on carbohydrate metabolism include polypeptides involved in sucrose or starch metabolism, carbon assimilation or carbohydrate transport, including, for example sucrose transporters or glucose/hexose transporters, enzymes involved in glycolysis/gluconeogenesis, the pentose phosphate cycle, or raffinose biosynthesis, and polypeptides involved in glucose signaling, such as SNF1 complex proteins.

Yield/Photosynthesis: Yield improvement resulting from increased photosynthesis. Polypeptides useful for increasing the rate of photosynthesis include phytochrome, photosystem I and II proteins, electron carriers, ATP synthase, NADH dehydrogenase and cytochrome oxidase.

5 Yield/Phosphorus: Yield improvement resulting from increased phosphorus uptake, transport or utilization. Polypeptides useful for improving yield in this manner include phosphatases and phosphate transporters.

Yield/Stress tolerance: Yield improvement resulting from improved plant growth and development by helping plants to tolerate stressful growth conditions. Polypeptides useful for improved stress tolerance under a variety of stress conditions include polypeptides involved in gene regulation, such as serine/threonine-protein kinases, MAP kinases, MAP kinase kinases, and MAP kinase kinase kinases; polypeptides that act as receptors for signal transduction and regulation, such as receptor protein kinases; intracellular signaling proteins, such as protein phosphatases, GTP binding proteins, and phospholipid signaling proteins; polypeptides involved in arginine biosynthesis; polypeptides involved in ATP metabolism, including for example ATPase, adenylate transporters, and polypeptides involved in ATP synthesis and transport; polypeptides involved in glycine betaine, jasmonic acid, flavonoid or steroid biosynthesis; and hemoglobin. Enhanced or reduced activity of such polypeptides in transgenic plants will provide changes in the ability of a plant to respond to a variety of environmental stresses, such as chemical stress, drought stress and pest stress.

Cold tolerance: Polypeptides of interest for improving plant tolerance to cold or freezing temperatures include polypeptides involved in biosynthesis of trehalose or raffinose, polypeptides encoded by cold induced genes, fatty acyl desaturases and other polypeptides

involved in glycerolipid or membrane lipid biosynthesis, which find use in modification of membrane fatty acid composition, alternative oxidase, calcium-dependent protein kinases, LEA proteins and uncoupling protein.

Heat tolerance: Polypeptides of interest for improving plant tolerance to heat include  
 5 polypeptides involved in biosynthesis of trehalose, polypeptides involved in glycerolipid biosynthesis or membrane lipid metabolism (for altering membrane fatty acid composition), heat shock proteins and mitochondrial NDK.

Osmotic tolerance: Polypeptides of interest for improving plant tolerance to extreme osmotic conditions include polypeptides involved in proline biosynthesis.

10 Drought tolerance: Polypeptides of interest for improving plant tolerance to drought conditions include aquaporins, polypeptides involved in biosynthesis of trehalose or wax, LEA proteins and invertase.

Pathogen or pest tolerance: Polypeptides of interest for improving plant tolerance to effects of plant pests or pathogens include proteases, polypeptides involved in anthocyanin  
 15 biosynthesis, polypeptides involved in cell wall metabolism, including cellulases, glucosidases, pectin methylesterase, pectinase, polygalacturonase, chitinase, chitosanase, and cellulose synthase, and polypeptides involved in biosynthesis of terpenoids or indole for production of bioactive metabolites to provide defense against herbivorous insects.

Cell cycle modification: Polypeptides encoding cell cycle enzymes and regulators of the  
 20 cell cycle pathway are useful for manipulating growth rate in plants to provide early vigor and accelerated maturation leading to improved yield. Improvements in quality traits, such as seed oil content, may also be obtained by expression of cell cycle enzymes and cell cycle regulators. Polypeptides of interest for modification of cell cycle pathway include cyclins and EIF5alpha

pathway proteins, polypeptides involved in polyamine metabolism, polypeptides which act as regulators of the cell cycle pathway, including cyclin-dependent kinases (CDKs), CDK-activating kinases, CDK-inhibitors, Rb and Rb-binding proteins, and transcription factors that activate genes involved in cell proliferation and division, such as the E2F family of transcription factors, proteins involved in degradation of cyclins, such as cullins, and plant homologs of tumor suppressor polypeptides.

Seed protein yield/content: Polypeptides useful for providing increased seed protein quantity and/or quality include polypeptides involved in the metabolism of amino acids in plants, particularly polypeptides involved in biosynthesis of methionine/cysteine and lysine, amino acid transporters, amino acid efflux carriers, seed storage proteins, proteases, and polypeptides involved in phytic acid metabolism.

Seed oil yield/content: Polypeptides useful for providing increased seed oil quantity and/or quality include polypeptides involved in fatty acid and glycerolipid biosynthesis, beta-oxidation enzymes, enzymes involved in biosynthesis of nutritional compounds, such as carotenoids and tocopherols, and polypeptides that increase embryo size or number or thickness of aleurone.

Disease response in plants: Polypeptides useful for imparting improved disease responses to plants include polypeptides encoded by cercosporin induced genes, antifungal proteins and proteins encoded by R-genes or SAR genes. Expression of such polypeptides in transgenic plants will provide an increase in disease resistance ability of plants.

Galactomannan biosynthesis: Polypeptides involved in production of galactomannans are of interest for providing plants having increased and/or modified reserve polysaccharides for use in food, pharmaceutical, cosmetic, paper and paint industries.

Flavonoid/isoflavonoid metabolism in plants: Polypeptides of interest for modification of flavonoid/isoflavonoid metabolism in plants include cinnamate-4-hydroxylase, chalcone synthase and flavonol synthase. Enhanced or reduced activity of such polypeptides in transgenic plants will provide changes in the quantity and/or speed of flavonoid metabolism in plants and  
 5 may improve disease resistance by enhancing synthesis of protective secondary metabolites or improving signaling pathways governing disease resistance.

Plant growth regulators: Polypeptides involved in production of substances that regulate the growth of various plant tissues are of interest in the present invention and may be used to provide transgenic plants having altered morphologies and improved plant growth and  
 10 development profiles leading to improvements in yield and stress response. Of particular interest are polypeptides involved in the biosynthesis of plant growth hormones, such as gibberellins, cytokinins, auxins, ethylene and abscisic acid, and other proteins involved in the activity and/or transport of such polypeptides, including for example, cytokinin oxidase, cytokinin/purine permeases, F-box proteins, G-proteins and phytosulfokines.

15 Herbicide tolerance: Polypeptides of interest for producing plants having tolerance to plant herbicides include polypeptides involved in the shikimate pathway, which are of interest for providing glyphosate tolerant plants. Such polypeptides include polypeptides involved in biosynthesis of chorismate, phenylalanine, tyrosine and tryptophan.

Transcription factors in plants: Transcription factors play a key role in plant growth and  
 20 development by controlling the expression of one or more genes in temporal, spatial and physiological specific patterns. Enhanced or reduced activity of such polypeptides in transgenic plants will provide significant changes in gene transcription patterns and provide a variety of beneficial effects in plant growth, development and response to environmental conditions.

Transcription factors of interest include, but are not limited to myb transcription factors, including helix-turn-helix proteins, homeodomain transcription factors, leucine zipper transcription factors, MADS transcription factors, transcription factors having AP2 domains, zinc finger transcription factors, CCAAT binding transcription factors, ethylene responsive  
5 transcription factors, transcription initiation factors and UV damaged DNA binding proteins.

Homologous recombination: Increasing the rate of homologous recombination in plants is useful for accelerating the introgression of transgenes into breeding varieties by backcrossing, and to enhance the conventional breeding process by allowing rare recombinants between closely linked genes in phase repulsion to be identified more easily. Polypeptides useful for  
10 expression in plants to provide increased homologous recombination include polypeptides involved in mitosis and/or meiosis, including for example, resolvases and polypeptide members of the RAD52 epistasis group.

Lignin biosynthesis: Polypeptides involved in lignin biosynthesis are of interest for increasing plants' resistance to lodging and for increasing the usefulness of plant materials as  
15 biofuels.

In one embodiment of the invention, the function of a recombinant polypeptide is determined by comparison of the amino acid sequence of the recombinant polypeptide to amino acid sequences of known polypeptides. A variety of homology based search algorithms are available to compare a query sequence to a protein database, including for example, BLAST,  
20 FASTA, and Smith-Waterman. In the present application, BLASTX and BLASTP algorithms are used to provide protein function information. A number of values are examined in order to assess the confidence of the function assignment. Useful measurements include "E-value" (also shown as "hit\_p"), "percent identity", "percent query coverage", and "percent hit coverage".

In BLAST, E-value, or expectation value, represents the number of different alignments with scores equivalent to or better than the raw alignment score,  $S$ , that are expected to occur in a database search by chance. The lower the E value, the more significant the match. Because database size is an element in E-value calculations, E-values obtained by BLASTing against public databases, such as GenBank, have generally increased over time for any given query/entry match. In setting criteria for confidence of polypeptide function prediction, a “high” BLAST match is considered herein as having an E-value for the top BLAST hit provided in Table 1 of less than  $1E-30$ ; a medium BLASTX E-value is  $1E-30$  to  $1E-8$ ; and a low BLASTX E-value is greater than  $1E-8$ . The top BLAST hit and corresponding E values are provided in Table 1.

Percent identity refers to the percentage of identically matched amino acid residues that exist along the length of that portion of the sequences which is aligned by the BLAST algorithm. In setting criteria for confidence of polypeptide function prediction, a “high” BLAST match is considered herein as having percent identity for the top BLAST hit provided in Table 1 of at least 70%; a medium percent identity value is 35% to 70%; and a low percent identity is less than 35%.

In one embodiment of the invention, the protein function assignment in the present invention is determined using combinations of E-values, percent identity, query coverage and hit coverage. Query coverage refers to the percent of the query sequence that is represented in the BLAST alignment. Hit coverage refers to the percent of the database entry that is represented in the BLAST alignment. In one embodiment of the invention,, function of a query polypeptide is inferred from function of a protein homolog where either (1)  $\text{hit\_p} < 1e-30$  or  $\% \text{ identity} > 35\%$  AND  $\text{query\_coverage} > 50\%$  AND  $\text{hit\_coverage} > 50\%$ , or (2)  $\text{hit\_p} < 1e-8$  AND  $\text{query\_coverage} > 70\%$  AND  $\text{hit\_coverage} > 70\%$ .



Another aspect of the invention comprises a functional variant which differs in one or more amino acids from those of a recombinant polypeptide provided herein as the result of one or more conservative amino acid substitutions. It is well known in the art that one or more amino acids in a reference sequence can be substituted with at least one other amino acid, the charge and polarity of which are similar to that of the native amino acid, resulting in a silent change. For instance, valine is a conservative substitute for alanine and threonine is a conservative substitute for serine. Conservative substitutions for an amino acid within a polypeptide sequence can be selected from other members of the class to which the naturally occurring amino acid belongs. Amino acids can be divided into the following four groups: (1) acidic amino acids, (2) basic amino acids, (3) neutral polar amino acids, and (4) neutral nonpolar amino acids. Representative amino acids within these various groups include, but are not limited to: (1) acidic (negatively charged) amino acids such as aspartic acid and glutamic acid; (2) basic (positively charged) amino acids such as arginine, histidine, and lysine; (3) neutral polar amino acids such as glycine, serine, threonine, cysteine, tyrosine, asparagine, and glutamine; and (4) neutral nonpolar (hydrophobic) amino acids such as alanine, leucine, isoleucine, valine, proline, phenylalanine, tryptophan, and methionine. Conserved substitutes for an amino acid within a polypeptide sequence can be selected from other members of the group to which the naturally occurring amino acid belongs. For example, a group of amino acids having aliphatic side chains is glycine, alanine, valine, leucine, and isoleucine; a group of amino acids having aliphatic-hydroxyl side chains is serine and threonine; a group of amino acids having amide-containing side chains is asparagine and glutamine; a group of amino acids having aromatic side chains is phenylalanine, tyrosine, and tryptophan; a group of amino acids having basic side chains is lysine, arginine, and histidine; and a group of amino acids having sulfur-containing side chains is cysteine and

methionine. Examples of conservative amino acid substitution groups are: valine-leucine, valine-isoleucine, phenylalanine-tyrosine, lysine-arginine, alanine-valine, aspartic acid-glutamic acid, and asparagine-glutamine. In one embodiment a recombinant polypeptide of the invention may differ in one or more amino acids as the result of deletion or insertion of one or more amino acids in a native sequence. See for example, US Patent number 5,500,365, which is hereby incorporated by reference.

One embodiment of the present invention is a variant which has the same function as a recombinant polypeptide provided herein, but with increased or decreased activity or altered specificity. Such variations in protein activity can be achieved by mutagenesis or may exist naturally in polypeptides encoded by related genes, for example in a related polypeptide encoded by a different allele or in a different species. Variant polypeptides may be obtained by well known nucleic acid or protein screening methods using DNA or antibody probes, for example by screening libraries for genes encoding related polypeptides, or in the case of expression libraries, by screening directly for variant polypeptides. Screening methods for obtaining a modified protein or enzymatic activity of interest by mutagenesis are disclosed in US Patent 5,939,250, which is hereby incorporated by reference. An alternative approach to the generation of variants uses random recombination techniques such as "DNA shuffling" as disclosed in US Patents 5,605,793; 5,811,238; 5,830,721 and 5,837,458; and International Applications WO 98/31837 and WO 99/65927, all of which are hereby incorporated by reference. An alternative method of molecular evolution involves a staggered extension process (StEP) for in vitro mutagenesis and recombination of nucleic acid molecule sequences, as disclosed in US Patent 5,965,408 and International Application WO 98/42832, both of which are hereby incorporated by reference.

Polypeptides of the present invention that are functional variants of the polypeptides provided herein will generally demonstrate significant identity with the polypeptides provided herein. One embodiment of the invention is a polypeptide having at least about 35% sequence identity, at least about 50% sequence identity, at least about 60% sequence identity, at least about 70% sequence identity, at least about 80% sequence identity, and at least about 85%, 90%, 95% or even greater sequence identity with a recombinant polypeptide sequence described herein. One embodiment of the invention is a polypeptide having an amino acid sequence provided herein (reference polypeptides) and functional variants of such reference polypeptide, wherein such functional variant comprises at least about 50 consecutive amino acids having at least about 90% identity to about a 50 amino acid polypeptide fragment of said reference polypeptide.

### **Recombinant DNA Constructs**

In one embodiment the invention encompasses the use of recombinant polynucleotides in recombinant constructs, i.e. constructs comprising recombinant polynucleotides that are constructed or modified outside of cells and that join nucleic acids that are not found joined in nature. Using methods known to those of ordinary skill in the art, recombinant polynucleotides of the invention can be inserted into recombinant DNA constructs that can then be introduced into a host cell of choice for expression of the encoded polypeptide or to provide for reduction of expression of the encoded polypeptide, for example by antisense or cosuppression methods. Potential host cells include both prokaryotic and eukaryotic cells. One embodiment of the invention uses a recombinant polynucleotide of the present invention for preparation of recombinant constructs for use in plant transformation.

In plant transformation, exogenous genetic material is transferred into a plant cell. As used herein “exogenous” refers to a nucleic acid molecule, for example a recombinant DNA construct comprising a recombinant polynucleotide of the present invention, produced outside the organism, e.g. plant, into which it is introduced. An exogenous nucleic acid molecule can have a naturally occurring or non-naturally occurring nucleic acid sequence. One skilled in the art recognizes that an exogenous nucleic acid molecule can be derived from the same species into which it is introduced or from a different species. Such exogenous genetic material may be transferred into either monocot or dicot plants including, but not limited to, soy, cotton, canola, maize, teosinte, wheat, rice, and *Arabidopsis* plants. Transformed plant cells comprising such exogenous genetic material may be regenerated to produce whole transformed plants.

Exogenous genetic material may be transferred into a plant cell by the use of a recombinant construct, also known as a vector, designed for such a purpose. A recombinant construct can comprise a number of sequence elements, including promoters, encoding regions, and selectable markers. Recombinant constructs are available which have been designed to replicate in both *E. coli* and *A. tumefaciens* and have all of the features required for transferring large inserts of DNA into plant chromosomes. Design of such vectors is generally within the skill of the art.

A recombinant construct will generally include a plant promoter to direct transcription of the protein-encoding region or the antisense sequence of choice. Numerous promoters, which are active in plant cells, have been described in the literature. These include the nopaline synthase (NOS) promoter and octopine synthase (OCS) promoters carried on tumor-inducing plasmids of *Agrobacterium tumefaciens* or caulimovirus promoters such as the *Cauliflower Mosaic Virus* (CaMV) 19S or 35S promoter (US Patent 5,352,605), and the *Figwort Mosaic*

*Virus* (FMV) 35S-promoter (US Patent 5,378,619). These promoters and numerous others have been used to create recombinant vectors for expression in plants. Any promoter known or found to cause transcription of DNA in plant cells can be used in the present invention. Other useful promoters are described, for example, in U.S. Patents No. 5,378,619; 5,391,725; 5,428,147; 5,447,858; 5,608,144; 5,614,399; 5,633,441; and 5,633,435, all of which are hereby incorporated by reference.

In addition, promoter enhancers, such as the CaMV 35S enhancer or a tissue specific enhancer, may be used to enhance gene transcription levels. Enhancers often are found 5' to the start of transcription in a promoter that functions in eukaryotic cells, but can often be inserted in the forward or reverse orientation 5' or 3' to the coding sequence. In some instances, these 5' enhancing elements are introns. Deemed to be particularly useful as enhancers are the 5' introns of the rice actin 1 and rice actin 2 genes. Examples of other enhancers which could be used in accordance with the invention include elements from octopine synthase genes, the maize alcohol dehydrogenase gene intron 1, elements from the maize shrunken 1 gene, the sucrose synthase intron, the TMV omega element, and promoters from non-plant eukaryotes.

Recombinant constructs can also contain one or more 5' non-translated leader sequences which serve to enhance polypeptide production from the resulting mRNA transcripts. Such sequences may be derived from the promoter selected to express the gene or can be specifically modified to increase translation of the mRNA. Such regions may also be obtained from viral RNAs, from suitable eukaryotic genes, or from a synthetic gene sequence. For a review of optimizing expression of transgenes, see Koziel *et al.* (1996) *Plant Mol. Biol.* 32:393-405).

Recombinant constructs may also include, with the coding region of interest, a nucleic acid sequence that acts, in whole or in part, to terminate transcription of that region. One type of

3' untranslated sequence which may be used is a 3' UTR from the nopaline synthase gene (nos 3') of *Agrobacterium tumefaciens*. Other 3' termination regions of interest include those from a gene encoding the small subunit of a ribulose-1,5-bisphosphate carboxylase-oxygenase (rbcS), and more specifically, from a rice rbcS gene (US Patent 6426446), the 3' UTR for the T7 transcript of *Agrobacterium tumefaciens*, the 3' end of the protease inhibitor I or II genes from potato or tomato, and the 3' region isolated from Cauliflower Mosaic Virus. Alternatively, one also could use a gamma coixin, oleosin 3 or other 3' UTRs from the genus *Coix* (PCT Publication WO 99/58659).

Recombinant constructs may also include a selectable marker. Selectable markers may be used to select for plants or plant cells that contain the exogenous genetic material. Useful selectable marker genes include those conferring resistance to antibiotics such as kanamycin (*nptII*), hygromycin B (*aph IV*) and gentamycin (*aac3* and *aacC4*) or resistance to herbicides such as glufosinate (*bar* or *pat*) and glyphosate (EPSPS). Examples of such selectable markers are illustrated in U.S. Patents 5,550,318; 5,633,435; 5,780,708 and 6,118,047, all of which are hereby incorporated by reference.

Recombinant constructs may also include a screenable marker. Screenable markers may be used to monitor transformation. Exemplary screenable markers include genes expressing a colored or fluorescent protein such as a luciferase or green fluorescent protein (GFP), a  $\beta$ -glucuronidase or uidA gene (GUS) which encodes an enzyme for which various chromogenic substrates are known or an R-locus gene, which encodes a product that regulates the production of anthocyanin pigments (red color) in plant tissues. Other possible selectable and/or screenable marker genes will be apparent to those of skill in the art.

Recombinant constructs may also include a transit peptide for targeting of a gene target to a plant organelle, particularly to a chloroplast, leucoplast or other plastid organelle, see for example U.S. Patent 5,188,642, which is hereby incorporated by reference.

For use in *Agrobacterium* mediated transformation methods, recombinant constructs of  
5 the present invention may also include T-DNA border regions flanking the DNA to be inserted into the plant genome to provide for transfer of the DNA into the plant host chromosome as discussed in more detail below. An exemplary plasmid that finds use in such transformation methods is pMON18365, a T-DNA vector that can be used to clone exogenous genes and transfer them into plants using *Agrobacterium*-mediated transformation. See published U.S.  
10 Patent Application 20030024014, which is hereby incorporated by reference. This vector contains the left border and right border sequences necessary for *Agrobacterium* transformation. The plasmid also has origins of replication for maintaining the plasmid in both *E. coli* and *Agrobacterium tumefaciens* strains.

A candidate gene is prepared for insertion into the T-DNA vector, for example using  
15 well-known gene cloning techniques such as PCR. Restriction sites may be introduced onto each end of the gene to facilitate cloning. For example, candidate genes may be amplified by PCR techniques using a set of primers. Both the amplified DNA and the cloning vector are cut with the same restriction enzymes, for example, *NotI* and *PstI*. The resulting fragments are gel-purified, ligated together, and transformed into *E. coli*. Plasmid DNA containing the vector with  
20 inserted gene may be isolated from *E. coli* cells selected for spectinomycin resistance, and the presence of the desired insert verified by digestion with the appropriate restriction enzymes. Undigested plasmid may then be transformed into *Agrobacterium tumefaciens* using techniques well known to those in the art, and transformed *Agrobacterium* cells containing the vector of

interest selected based on spectinomycin resistance. These and other similar recombinant constructs useful for plant transformation may be readily prepared by one skilled in the art.

### **Transformation Methods and Transgenic Plants**

Methods and compositions for transforming bacteria and other microorganisms are known in the art. See for example *Molecular Cloning: A Laboratory Manual, 3<sup>rd</sup> edition Volumes 1, 2, and 3*. J.F. Sambrook, D.W. Russell, and N. Irwin, Cold Spring Harbor Laboratory Press, 2000.

Technology for introduction of DNA into cells is well known to those of skill in the art. Methods and materials for transforming plants by introducing a transgenic DNA construct into a plant genome in the practice of this invention can include any of the well-known and demonstrated methods including electroporation as illustrated in U.S. Patent 5,384,253, microprojectile bombardment as illustrated in U.S. Patents 5,015,580; 5,550,318; 5,538,880; 6,160,208; 6,399,861 and 6,403,865, Agrobacterium-mediated transformation as illustrated in U.S. Patents 5,635,055; 5,824,877; 5,591,616; 5,981,840 and 6,384,301, and protoplast transformation as illustrated in U.S. Patents 5,508,184, all of which are hereby incorporated by reference.

Any of the recombinant polynucleotides of the present invention may be introduced into a plant cell in a permanent or transient manner in combination with other genetic elements such as vectors, promoters enhancers etc. Further any of the recombinant polynucleotides of the present invention may be introduced into a plant cell in a manner that allows for production of the polypeptide or fragment thereof encoded by the recombinant polynucleotide in the plant cell, or in a manner that provides for decreased expression of an endogenous gene and concomitant decreased production of protein.



It is also to be understood that two different transgenic plants can also be mated to produce offspring that contain two independently segregating added, exogenous genes. Selfing of appropriate progeny can produce plants that are homozygous for both added, exogenous genes that encode a polypeptide of interest. Back-crossing to a parental plant and out-crossing with a non-transgenic plant are also contemplated, as is vegetative propagation.

Expression of the recombinant polynucleotides of the present invention and the concomitant production of polypeptides encoded by the recombinant polynucleotides is of interest for production of transgenic plants having improved properties, particularly, improved properties which result in crop plant yield improvement. Expression of recombinant polypeptides of the present invention in plant cells may be evaluated by specifically identifying the protein products of the introduced genes or evaluating the phenotypic changes brought about by their expression. It is noted that when the polypeptide being produced in a transgenic plant is native to the target plant species, quantitative analyses comparing the transformed plant to wild type plants may be required to demonstrate increased expression of the polypeptide of this invention.

Assays for the production and identification of specific proteins make use of various physical-chemical, structural, functional, or other properties of the proteins. Unique physical-chemical or structural properties allow the proteins to be separated and identified by electrophoretic procedures, such as native or denaturing gel electrophoresis or isoelectric focusing, or by chromatographic techniques such as ion exchange or gel exclusion chromatography. The unique structures of individual proteins offer opportunities for use of specific antibodies to detect their presence in formats such as an ELISA assay. Combinations of approaches may be employed with even greater specificity such as western blotting in which

antibodies are used to locate individual gene products that have been separated by electrophoretic techniques. Additional techniques may be employed to absolutely confirm the identity of the product of interest such as evaluation by amino acid sequencing following purification. Although these are among the most commonly employed, other procedures may be additionally used.

Assay procedures may also be used to identify the expression of proteins by their functionality, particularly where the expressed protein is an enzyme capable of catalyzing chemical reactions involving specific substrates and products. These reactions may be measured, for example in plant extracts, by providing and quantifying the loss of substrates or the generation of products of the reactions by physical and/or chemical procedures.

In many cases, the expression of a gene product is determined by evaluating the phenotypic results of its expression. Such evaluations may be simply as visual observations, or may involve assays. Such assays may take many forms including but not limited to analyzing changes in the chemical composition, morphology, or physiological properties of the plant. Chemical composition may be altered by expression of genes encoding enzymes or storage proteins which change amino acid composition and may be detected by amino acid analysis, or by enzymes which change starch quantity which may be analyzed by near infrared reflectance spectrometry. Morphological changes may include greater stature or thicker stalks.

Plants with decreased expression of a gene of interest can also be achieved through the use of polynucleotides of the present invention, for example by expression of antisense nucleic acids, or by identification of plants transformed with sense expression constructs that exhibit cosuppression effects.

Antisense approaches are a way of preventing or reducing gene function by targeting the genetic material as disclosed in U.S. Patents 4,801,540; 5,107,065; 5,759,829; 5,910,444; 6,184,439; and 6,198,026, all of which are hereby incorporated by reference. The objective of the antisense approach is to use a sequence complementary to the target gene to block its expression and create a mutant cell line or organism in which the level of a single chosen protein is selectively reduced or abolished. Antisense techniques have several advantages over other 'reverse genetic' approaches. The site of inactivation and its developmental effect can be manipulated by the choice of promoter for antisense genes or by the timing of external application or microinjection. Antisense can manipulate its specificity by selecting either unique regions of the target gene or regions where it shares homology to other related genes.

The principle of regulation by antisense RNA is that RNA that is complementary to the target mRNA is introduced into cells, resulting in specific RNA:RNA duplexes being formed by base pairing between the antisense substrate and the target. Under one embodiment, the process involves the introduction and expression of an antisense gene sequence. Such a sequence is one in which part or all of the normal gene sequences are placed under a promoter in inverted orientation so that the 'wrong' or complementary strand is transcribed into a noncoding antisense RNA that hybridizes with the target mRNA and interferes with its expression. An antisense vector is constructed by standard procedures and introduced into cells by transformation, transfection, electroporation, microinjection, infection, etc. The type of transformation and choice of vector will determine whether expression is transient or stable. The promoter used for the antisense gene may influence the level, timing, tissue, specificity, or inducibility of the antisense inhibition.

As used herein “gene suppression” means any of the well-known methods for suppressing expression of protein from a gene including sense suppression, anti-sense suppression and RNAi suppression. In suppressing genes to provide plants with a desirable phenotype, anti-sense and RNAi gene suppression methods are preferred. For a description of anti-sense regulation of gene expression in plant cells see U.S. Patent 5,107,065. For a description of RNAi gene suppression in plants by transcription of a dsRNA see U.S. Patent 6,506,559, U.S. Patent Application Publication No. 2002/0168707 A1, and U.S. Patent Applications Serial No. 09/423,143 (see WO 98/53083), 09/127,735 (see WO 99/53050), and 09/084,942 (see WO 99/61631), all of which are hereby incorporated by reference. Suppression of an gene by RNAi can be achieved using a recombinant DNA construct having a promoter operably linked to a DNA element comprising a sense and anti-sense element of a segment of genomic DNA of the gene, e.g., a segment of at least about 23 nucleotides, more preferably about 50 to 200 nucleotides where the sense and anti-sense DNA components can be directly linked or joined by an intron or artificial DNA segment that can form a loop when the transcribed RNA hybridizes to form a hairpin structure. For example, genomic DNA from a polymorphic locus of SEQ ID NO: 1 through SEQ ID NO: 31,564 can be used in a recombinant construct for suppression of a cognate gene by RNAi suppression.

Insertion mutations created by transposable elements may also prevent gene function. For example, in many dicot plants, transformation with the T-DNA of *Agrobacterium* may be readily achieved and large numbers of transformants can be rapidly obtained. Also, some species have lines with active transposable elements that can efficiently be used for the generation of large numbers of insertion mutations, while some other species lack such options. Mutant plants produced by *Agrobacterium* or transposon mutagenesis and having altered

expression of a polypeptide of interest can be identified using the polynucleotides of the present invention. For example, a large population of mutated plants may be screened with polynucleotides encoding the polypeptide of interest to detect mutated plants having an insertion in the gene encoding the polypeptide of interest.

5 In one embodiment of the invention, polynucleotides of the present invention may be used in site-directed mutagenesis. Site-directed mutagenesis may be utilized to modify nucleic acid sequences, particularly as it is a technique that allows one or more of the amino acids encoded by a nucleic acid molecule to be altered (*e.g.*, a threonine to be replaced by a methionine). Three basic methods for site-directed mutagenesis are often employed. These are  
10 cassette mutagenesis, primer extension, and methods based upon PCR.

In addition to the above discussed procedures, practitioners are familiar with the standard resource materials which describe specific conditions and procedures for the construction, manipulation and isolation of macromolecules (*e.g.*, DNA molecules, plasmids, etc.), generation of recombinant organisms and the screening and isolating of clones.

## 15 Arrays

In one embodiment of the invention, the recombinant polynucleotides or recombinant polypeptides of this invention may be used to prepare arrays of target molecules arranged on a surface of a substrate. The target molecules may be known molecules, *e.g.* polynucleotides (including oligonucleotides) or polypeptides, which are capable of binding to specific probes,  
20 such as complementary nucleic acids or specific antibodies. The target molecules may be immobilized, *e.g.* by covalent or non-covalent bonding, to the surface in small amounts of substantially purified and isolated molecules in a grid pattern. By immobilized it is meant that the target molecules maintain their position relative to the solid support under hybridization and

washing conditions. Target molecules are deposited in small footprint, isolated quantities of “spotted elements” of preferably single-stranded polynucleotide preferably arranged in rectangular grids in a density of about 30 to 100 or more, e.g. up to about 1000, spotted elements per square centimeter. In one embodiment of the invention, the arrays comprise at least about 5 100 or more, e.g. at least about 1000 to 5000, distinct target polynucleotides per unit substrate. Where detection of transcription for a large number of genes is desired, the economics of arrays favors a high density design criteria provided that the target molecules are sufficiently separated so that the intensity of the indicia of a binding event associated with highly expressed probe molecules does not overwhelm and mask the indicia of neighboring binding events. For high- 10 density microarrays each spotted element may contain up to about  $10^7$  or more copies of the target molecule, e.g. single stranded cDNA, on glass substrates or nylon substrates.

Arrays of this invention may be prepared with molecules from a single species, preferably a plant species, or with molecules from other species, particularly other plant species. Arrays with target molecules from a single species can be used with probe molecules from the 15 same species or a different species due to the ability of cross species homologous genes to hybridize. It is generally preferred for high stringency hybridization that the target and probe molecules are from the same species.

In one embodiment of the invention, the organism of interest is a plant and the target molecules are polynucleotides or oligonucleotides with nucleic acid sequences having at least 20 about 80 percent sequence identity to a corresponding sequence of the same length in a recombinant polynucleotide having a sequence selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 31,564 or complements thereof. In another embodiment of the invention, at least about 10% of the target molecules on an array have at least about 15

consecutive nucleotides of sequence having at least about 80% and up to about 100% identity with a corresponding sequence of the same length in a recombinant polynucleotide having a sequence selected from the group consisting of SEQ ID NO: 1 through SEQ ID NO: 31,564 or complements or fragments thereof.

5        Such arrays are useful in a variety of applications, including gene discovery, genomic research, molecular breeding and bioactive compound screening. One use of arrays is in the analysis of differential gene transcription, e.g. transcription profiling where the production of mRNA in different cells, normally a cell of interest and a control, is compared and discrepancies in gene expression are identified. In such assays, the presence of discrepancies indicates a  
10 difference in gene expression levels in the cells being compared. Such information is useful for the identification of the types of genes expressed in a particular cell or tissue type in a known environment. Such applications generally involve the following steps: (a) preparation of probe, e.g. attaching a label to a plurality of expressed molecules; (b) contact of probe with the array under conditions sufficient for probe to bind with corresponding target, e.g. by hybridization or  
15 specific binding; (c) removal of unbound probe from the array; and (d) detection of bound probe.

A probe may be prepared with RNA extracted from a given cell line or tissue. The probe may be produced by reverse transcription of mRNA or total RNA and labeled with radioactive or fluorescent labeling. A probe is typically a mixture containing many different sequences in various amounts, corresponding to the numbers of copies of the original mRNA species  
20 extracted from the sample.

The initial RNA sample for probe preparation will typically be derived from a physiological source. The physiological source may be selected from a variety of organisms, with physiological sources of interest including single celled organisms such as yeast and

multicellular organisms, including plants and animals, particularly plants, where the physiological sources from multicellular organisms may be derived from particular organs or tissues of the multicellular organism, or from isolated cells derived from an organ, or tissue of the organism. The physiological sources may also be multicellular organisms at different developmental stages (*e.g.*, 10-day-old seedlings), or organisms grown under different environmental conditions (*e.g.*, drought-stressed plants) or treated with chemicals.

In preparing the RNA probe, the physiological source may be subjected to a number of different processing steps, where such processing steps might include tissue homogenation, cell isolation and cytoplasmic extraction, nucleic acid extraction and the like, where such processing steps are known to those of skill in the art. Methods of isolating RNA from cells, tissues, organs or whole organisms are well known to those skilled in the art.

#### **Computer Based Systems and Methods**

In one embodiment of the invention, the sequence of the molecules of this invention can be provided in a variety of media to facilitate use thereof. Such media may provide a subset thereof in a form that allows a skilled artisan to examine the sequences. In a one embodiment, about 20, about 50, about 100, and about 200 or more of the polynucleotide and/or the polypeptide sequences of the present invention can be recorded on computer readable media. As used herein, "computer readable media" refers to any medium that can be read and accessed directly by a computer. Such media include, but are not limited to: magnetic storage media, such as floppy discs, hard disc, storage medium, and magnetic tape; optical storage media such as CD-ROM; electrical storage media such as RAM and ROM; and hybrids of these categories such as magnetic/optical storage media. A skilled artisan can readily appreciate how any of the presently known computer readable media can be used to create a manufacture comprising a



computer readable medium having recorded thereon a nucleotide sequence of the present invention.

As used herein, "recorded" refers to a process for storing information on computer readable media. A skilled artisan can readily adopt any of the presently known methods for recording information on computer readable media to generate media comprising the nucleotide sequence information of the present invention. A variety of data storage structures are available to a skilled artisan for creating a computer readable medium having recorded thereon a nucleotide sequence of the present invention. The choice of the data storage structure will generally be based on the means chosen to access the stored information. In addition, a variety of data processor programs and formats can be used to store the nucleotide sequence information of the present invention on computer readable media. The sequence information can be represented in a word processing text file, formatted in commercially-available software such as WordPerfect and Microsoft Word, or represented in the form of an ASCII file, stored in a database application, such as DB2, Sybase, Oracle, or the like. A skilled artisan can readily adapt any number of data processor structuring formats (*e.g.*, text file or database) in order to obtain a computer readable medium having recorded thereon the nucleotide sequence information of the present invention.

By providing one or more of polynucleotide or polypeptide sequences of the present invention in a computer readable medium, a skilled artisan can routinely access the sequence information for a variety of purposes. The examples which follow demonstrate how software which implements the BLAST and BLAZE search algorithms on a Sybase system can be used to identify open reading frames (ORFs) within the genome that contain homology to ORFs or polypeptides from other organisms. Such ORFs are polypeptide encoding fragments within the

sequences of the present invention and are useful in producing commercially important polypeptides such as enzymes used in amino acid biosynthesis, metabolism, transcription, translation, RNA processing, nucleic acid and a protein degradation, protein modification, and DNA replication, restriction, modification, recombination, and repair.

5           One embodiment of the invention provides systems, particularly computer-based systems, which contain the sequence information described herein. Such systems are designed to identify commercially important fragments of the nucleic acid molecule of the present invention. As used herein, “a computer-based system” refers to the hardware, software, and memory used to analyze the sequence information of the present invention. A skilled artisan can readily  
10 appreciate that any one of the currently available computer-based systems are suitable for use in the present invention.

As indicated above, the computer-based systems of the present invention comprise a database having stored therein a polynucleotide sequence, polypeptide sequence, or both of the present invention and the necessary hardware and software for supporting and implementing a  
15 homology search. As used herein, “database” refers to memory system that can store searchable nucleotide sequence information. As used herein “query sequence” is a polynucleotide sequence, or a polypeptide sequence, or a polynucleotide sequence corresponding to a polypeptide sequence, or a polypeptide sequence corresponding to a polynucleotide sequence, that is used to query a collection of polynucleotide or polypeptide sequences. As used herein,  
20 “homology search” refers to one or more programs which are implemented on the computer-based system to compare a query sequence, *i.e.*, gene or peptide or a conserved region (motif), with the sequence information stored within the database. Homology searches are used to identify segments and/or regions of the sequence of the present invention that match a particular

query sequence. A variety of known searching algorithms are incorporated into commercially available software for conducting homology searches of databases and computer readable media comprising sequences of molecules of the present invention.

Sequence length of a query sequence may be from about 10 to about 100 or more amino acid residues or from about 20 to about 300 or more nucleotide residues. There are a variety of motifs known in the art. Protein motifs include, but are not limited to, enzymatic active sites and signal sequences. An amino acid query is converted to all of the nucleic acid sequences that encode that amino acid sequence by a software program, such as TBLASTN, which is then used to search the database. Nucleic acid query sequences that are motifs include, but are not limited to, promoter sequences, cis elements, hairpin structures and inducible expression elements (protein binding sequences).

One embodiment of the invention, provides an input device for receiving a query sequence, a memory for storing sequences (the query sequences of the present invention and sequences identified using a homology search as described above), and an output device for outputting the identified homologous sequences. A variety of structural formats for the input and output presentations can be used to input and output information in the computer-based systems of the present invention. One format for an output presentation ranks fragments of the sequence of the present invention by varying degrees of homology to the query sequence. Such presentation provides a skilled artisan with a ranking of sequences that contain various amounts of the query sequence and identifies the degree of homology contained in the identified fragment.

Having now generally described the invention, the same will be more readily understood through reference to the following examples which are provided by way of illustration, and are not intended to be limiting of the present invention, unless specified.

### Example 1

5           A cDNA library is generated from *Sorghum* tissue. Tissue is harvested and immediately frozen in liquid nitrogen. The harvested tissue is stored at -80°C until preparation of total RNA. The total RNA is purified using Trizol reagent from Invitrogen Corporation (Invitrogen Corporation, Carlsbad, California, U.S.A.), essentially as recommended by the manufacturer. Poly A<sup>+</sup> RNA (mRNA) is purified using magnetic oligo dT beads essentially as recommended  
10 by the manufacturer (Dynabeads, Dynal Biotech, Oslow, Norway).

Construction of plant cDNA libraries is well known in the art and a number of cloning strategies exist. A number of cDNA library construction kits are commercially available. cDNA libraries are prepared using the Superscript™ Plasmid System for cDNA synthesis and Plasmid Cloning (Invitrogen Corporation, Carlsbad, California, U.S.A.), as described in the Superscript II  
15 cDNA library synthesis protocol. The cDNA libraries are quality controlled for a good insert:vector ratio.

The cDNA libraries are plated on LB agar containing the appropriate antibiotics for selection and incubated at 37° for a sufficient time to allow the growth of individual colonies. Single colonies are individually placed in each well of a 96-well microtiter plates containing LB  
20 liquid including the selective antibiotics. The plates are incubated overnight at approximately 37°C with gentle shaking to promote growth of the cultures. The plasmid DNA is isolated from

each clone using Qiaprep plasmid isolation kits, using the conditions recommended by the manufacturer (Qiagen Inc., Valencia, California U.S.A.).

The template plasmid DNA clones are used for subsequent sequencing. Sequences of recombinant polynucleotides may be obtained by a number of sequencing techniques known in the art, including fluorescence-based sequencing methodologies. These methods have the detection, automation, and instrumentation capability necessary for the analysis of large volumes of sequence data. With these types of automated systems, fluorescent dye-labeled sequence reaction products are detected and data entered directly into the computer, producing a chromatogram that is subsequently viewed, stored, and analyzed using the corresponding software programs. These methods are known to those of skill in the art and have been described and reviewed.

## Example 2

The open reading frame in each recombinant polynucleotide sequence is identified by a combination of predictive and homology based methods. The longest open reading frame (ORF) is determined, and the top BLAST match is identified by BLASTX against NCBI. The top BLAST hit is then compared to the predicted ORF, with the BLAST hit given precedence in the case of discrepancies.

Functions of polypeptides encoded by the polynucleotide sequences of the present invention are determined using a hierarchical classification tool, termed FunCAT, for Functional Categories Annotation Tool. Most categories collected in FunCAT are classified by function, although other criteria are used, for example, cellular localization or temporal process. The assignment of a functional category to a query sequence is based on BLASTX sequence search results, which compare two protein sequences. FunCAT assigns categories by iteratively

scanning through all blast hits, starting with the most significant match, and reporting the first category assignment for each FunCAT source classification scheme. In the present invention, function of a query polypeptide is inferred from the function of a protein homolog where either

(1)  $\text{hit\_p} < 1\text{e-}30$  or  $\% \text{ identity} > 35\%$  AND  $\text{query\_coverage} > 50\%$  AND  $\text{hit\_coverage} > 50\%$ ,

5 or (2)  $\text{hit\_p} < 1\text{e-}8$  AND  $\text{query\_coverage} > 70\%$  AND  $\text{hit\_coverage} > 70\%$ .

Functional assignments from five public classification schemes, GO\_BP, GO\_CC, GO\_MF, KEGG, and EC, and one internal Monsanto classification scheme, POI, are provided in Table 1. The column under the heading “CAT\_TYPE” indicates the source of the classification. GO\_BP = Gene Ontology Consortium - biological process; GO\_CC = Gene Ontology Consortium - cellular component; GO\_MF = Gene Ontology Consortium - molecular function;

10 KEGG = KEGG functional hierarchy; EC = Enzyme Classification from ENZYME data bank release 25.0; POI = Pathways of Interest. The column under the heading “CAT\_DESC” provides the name of the subcategory into which the query sequence was classified. The column under the heading “PRODUCT\_HIT\_DESC” provides a description of the BLAST hit to the query

15 sequences that led to the specific classification. The column under the heading “HIT\_E” provides the e-value for the BLAST hit. It is noted that the e-value in the HIT\_E column may differ from the e-value based on the top BLAST hit provided in the E\_VALUE column since these calculations were done on different days, and database size is an element in E-value calculations. E-values obtained by BLASTing against public databases, such as GenBank, will

20 generally increase over time for any given query/entry match.

Sequences useful for producing transgenic plants having improved biological properties are identified from their FunCAT annotations and are also provided in Table 1. A biological property of particular interest is plant yield. Plant yield may be improved by alteration of a

variety of plant pathways, including those involving nitrogen, carbohydrate, or phosphorus utilization and/or uptake. Plant yield may also be improved by alteration of a plant's photosynthetic capacity or by improving a plant's ability to tolerate a variety of environmental stresses, including cold, heat, drought and osmotic stresses. Other biological properties of interest that may be improved using sequences of the present invention include pathogen or pest tolerance, herbicide tolerance, disease resistance, growth rate (for example by modification of cell cycle, by expression of transcription factors, or expression of growth regulators), seed oil and/or protein yield and quality, rate and control of recombination, and lignin content.

Sequences of recombinant polynucleotides are provided herein as SEQ ID NO: 1 through SEQ ID NO: 31,564 and sequences of recombinant polypeptides are provided as SEQ ID NO: 31,565 through SEQ ID NO: 63,128. Descriptions of each of these recombinant polynucleotide and recombinant polypeptide sequences are provided in Table 1.

#### **Table 1 Column Descriptions**

**SEQ\_NUM** provides the SEQ ID NO for the listed recombinant polynucleotide sequences.

**CONTIG\_ID** provides an arbitrary sequence name taken from the name of the clone from which the cDNA sequence was obtained.

**PROTEIN\_NUM** provides the SEQ ID NO for the recombinant polypeptide sequence

**NCBI\_GI** provides the GenBank ID number for the top BLAST hit for the sequence. The top BLAST hit is indicated by the National Center for Biotechnology Information GenBank Identifier number.

**NCBI\_GI\_DESCRIPTION** refers to the description of the GenBank top BLAST hit for the sequence.

**E\_VALUE** provides the expectation value for the top BLAST match.

**MATCH\_LENGTH** provides the length of the sequence which is aligned in the top BLAST match

**TOP\_HIT\_PCT\_IDENT** refers to the percentage of identically matched nucleotides (or residues) that exist along the length of that portion of the sequences which is aligned in the top BLAST match.

**CAT\_TYPE** indicates the classification scheme used to classify the sequence. **GO\_BP** = Gene Ontology Consortium - biological process; **GO\_CC** = Gene Ontology Consortium - cellular component; **GO\_MF** = Gene Ontology Consortium - molecular function; **KEGG** = KEGG functional hierarchy (KEGG = Kyoto Encyclopedia of Genes and Genomes); **EC** = Enzyme Classification from ENZYME data bank release 25.0; **POI** = Pathways of Interest.

**CAT\_DESC** provides the classification scheme subcategory to which the query sequence was assigned.

**PRODUCT\_CAT\_DESC** provides the FunCAT annotation category to which the query sequence was assigned.

**PRODUCT\_HIT\_DESC** provides the description of the BLAST hit which resulted in assignment of the sequence to the function category provided in the **cat\_desc** column.

**HIT\_E** provides the E value for the BLAST hit in the **hit\_desc** column.

**PCT\_IDENT** refers to the percentage of identically matched nucleotides (or residues) that exist along the length of that portion of the sequences which is aligned in the BLAST match provided in **hit\_desc**.

**QRY\_RANGE** lists the range of the query sequence aligned with the hit.

**HIT\_RANGE** lists the range of the hit sequence aligned with the query.



**QRY\_CVRG** provides the percent of query sequence length that matches to the hit (NCBI) sequence in the BLAST match ( $\% \text{ qry cvrg} = (\text{match length} / \text{query total length}) \times 100$ ).

**HIT\_CVRG** provides the percent of hit sequence length that matches to the query sequence in the match generated using BLAST ( $\% \text{ hit cvrg} = (\text{match length} / \text{hit total length}) \times 100$ ).

5

All publications and patent applications cited herein are hereby incorporated by reference in their entirety to the same extent as if each individual publication or patent application was specifically and individually indicated to be incorporated by reference.

Although the foregoing invention has been described in some detail by way of illustration and example for purposes of clarity of understanding, it will be obvious that certain changes and modifications may be practiced within the scope of the appended claims.

10